White Paper Report

Report ID: 98421

Application Number: HJ5002810

Project Director: William Thomas (wthomas4@unlnotes01.unl.edu)

Institution: University of Nebraska, Board of Regents

Reporting Period: 1/1/2010-3/31/2011

Report Due: 6/30/2011

Date Submitted: 9/12/2011

White Paper
Digging into Data Award
National Endowment for the Humanities
April 15, 2011

Railroads and the Making of Modern America: Tools for Spatio-temporal Correlation, Analysis, and Visualization

PIs: William G. Thomas, University of Nebraska, and Richard Healey, University of Portsmouth

Big History, Big Data, and Continental-Scale Geography:

A Report on Digging into Railroad Data

Overview

The railroad time table was an astonishingly complex innovation and perhaps the best example we have of an "interface" to the railroad system. At first, time tables were printed on small cards because local railroads, such as the Boston & Worcester, made just one or two runs a day over a short distance. But the 1850s marked a major shift as hundreds of new junctions came on line in the Midwest and South. Railroads in the 1850s operated primarily as passenger lines, and time tables for longer lines became increasingly intricate. The time table presented Americans with an abstract geography, a chart of time, cost, place, and distance.

But the nineteenth-century time table has proven very difficult to represent, even with computer technology, and especially difficult to reconstitute. This is true because of two essential, but often overlooked, phenomena: 1.) the railroad network was always changing, never the same at any given time, scale, or place, though people acted as if it were and 2.) when people read the time table, they did so with their own experience and information at hand, traversing a geographic scale from the local to the regional to the national.

The railroad, like the Internet, was both a hardware system and a set of software interfaces, both of which were socially constructed and applied with equal energy to a wide range of uses. Between 1850 and 1900 the railroad network of the United States expanded from a total length of about 5,000 miles to almost 200,000 miles. Growth was uneven across time, but the 1850s stands out as an important burst, when over 22,000 miles was constructed in the Midwest and South. Railroads touched nearly every aspect of modern society. The companies participated as active agents across an immense variety of social, environmental, and landscape processes: racial segregation, the settlement of coal mining communities, tourism, worker mobility, immigration, political corruption, land sales, the destruction of the bison, forest destruction and conservation, and irrigation.

Nineteenth-century Americans, in short, experienced something similar to our circumstances today: a major technological transformation in which temporal and spatial relations were reconfigured, a "second nature" system constructed, and personal mobility and identity altered. In this respect the initial burst of the railroad and telegraph era, from about 1840 to 1870, was one of the first transformative technological periods in American history. The nineteenth-century expression for this effect was to say that railroads "annihilated space and time." But, of course, railroads did not eradicate space and time; instead, they reconstituted these relationships.[1]

People interacted with the railroad as a technology system through gateways of perception and understanding that were both railroad constructed and socially constructed. The timetable offered one "interface" to the railroad, but there were many others, and any account of the railroad's effects would need to include newspapers, publications, print materials, rate tables, payrolls, lithographs, photographs, and annual reports. Newspapers reported on railroad developments planned, under construction, and completed. Timetables revealed a network of variable mobility and spatial possibility that was uneven. System maps provided selected spatial information about the network. Payrolls documented the location of workers along a set of changing divisions. Company annual reports described detailed information about the network's spatial and temporal configuration. The railroad

network, then, varied widely across space and time.[2]

In the field of Digital Humanities, many projects in digital history have proceeded to create "intentional archives" or "thematic research archives"--around a year, such as 1896 or around a place, such as Frank Zephyr's The Terrain of History (Rio de Jinero) or the Virtual Shanghai project. The digital articles published by the American Historical Review represent some of the most advanced experiments in historical form--in hyper-textual history. But in each case historians have, understandably, held either place or time relatively constant, in service of an effort to understand the other more deeply. As historians have taken "the spatial turn," they have also begun to explore "longue durée" approaches to specific places. Philip J. Ethington's Hypercities project and, in particular, Ghost Metropolis: Los Angeles Since 13,000 BP demonstrate the extraordinary complexity of integrating temporal and spatial information.[3]

Fernand Braudel, the great historian and founder of the Annales School, wrote in the late 1950s about the difficulty of the historian's quest to knit together time, place, and evidence: "a historian . . . will always wish to grasp the whole, the totality of social life. So he is led to bring together different levels, time spans, different kinds of time, structures, conjunctures, events." For Braudel social reality was "at every instant, synchronous with its history, a constantly changing image, although it might repeat a thousand previous details of a thousand previous realities." The goal of integrating social experience at different scales has been with us for some time of course, and so has the question of narrative form. Hugh Trevor-Roper asked decades ago, "How can one *both* move *and* carry along with one the fermenting depths which are also, at every point, influenced by the pressure of events around them? And how can one possibly do this so that the result is readable? That is the problem."[4]

More recently, Edward L. Ayers has argued "for the intricate interplay of the structural and the ephemeral, the enduring and the emergent. This is 'deep contingency,' a view of social life that fuses an active sense of place and an active sense of time." Ayers and his colleagues at the University of Richmond are creating "cinematic maps" to capture the voting patterns of Americans over two hundred

years and reveal patterns of social change at multiple levels.[5]

What historians and geographers want to do is no less than to integrate what have been macro and micro historical approaches, and to satisfy our longing to join together, on the one hand, the deep currents of social change we recognize as "climactic" or "tidal" in their sweep and on the other hand what we might constitute as another dimension of historical change and experience, the individual and communal.[6] One influential school of social theory has stressed "practice theory" to resolve these different scales of human experience. Anthony Giddens, Pierre Bourdieu, Marshall Sahlins, Raymond Williams, and Sherry Ortner, and others suggest that people and psychological processes are embedded in and inseparable from their physical and social contexts. These scholars have set aside older Marxian, neo-classical, and structuralist models, instead emphasizing that the material and the cultural have always been part of the same interrelated process of social change.[7]

Another approach to the problem, growing out of the field of science and technology studies, has developed around actor-network theory. Bruno Latour, in his landmark exploration of modernity We Have Never Been Modern, has argued for the close analysis of what he calls "translations" or "mediators," stressing the continuous process of exchange between nonhuman and human domains, a recognition foreclosed by the "modern constitution." With Actor-Network Theory, the method for mapping or revealing these mediators emphasizes flattening time and space in the analysis of evidence. Follow the actors, Latour tells us, and rebuild their network of associations. Only then can we move beyond the (false) dichotomy of macro and micro levels of analysis. We have privileged the separation of the human and natural for too long, and in so doing we seek to explain social action in a way that assumes *a priori* an exogenous force on a social system. For Latour all points on a network perform both global and local functions--they translate, mediate, and extend the effects of the interaction.[8]

When we examine evidence about the railroad in newspapers, annual reports, or other sources, we see the various scales at which the railroads operated and both the intensity and extensiveness of the process of transformation that accompanied the railroads. We are confronted with a data problem and a

representation problem. Articles, for example, do not provide systematic treatment of topics such as railroad development over time. Rather, they present, in almost random order, tiny and incomplete pieces of a much larger dynamic picture. Therefore they need to be placed within a broader multi-dimensional information infrastructure that provides essential context and allows the significance of the individual text fragments to be assessed. This problem is not related only to railroad history, but to the wider issue of studying "second nature" systems of any kind in the current digital humanities field.[9]

In selecting such a geographically uneven, socially complex, and widely experienced process as railroad development, we quickly realize that the process was highly dynamic and contingent. Places gained prominence and proximity through the rate and time structures that railroads imposed; people used the emerging network to advance time through mobility and travel. Transformations in one place (a new wharf, a lower rate, a faster service) rippled out affecting communities in another. The railroad network, however fixed with spikes and tracks, was never stable; its geography was ever shifting and recursive. We need better tools to visualize the second nature systems that railroads fostered if we are going to assess their social consequences.

We seek to use the digital tools available today help us examine and reconstitute the temporal and spatial landscape so that we might see the relations among social actors better. We also seek to create different forms of scholarly communication for integrating time, place, and evidence. If we focus less on causation and more on process and consequences, on how events resonated in society, the digital environment may help us understand broad transformations at different scales. The ACLS Cyberinfrastructure Report issued in 2005 noted that "humanists and social scientists and their organizations must build the tools and standards they need: others will not do it for them." And it pointed to the new ways of seeing and knowing that digital humanities can afford. A whole generation of digital tools and techniques might offer alternative forms for historical scholarship and alternative models of social change.[10]

We attempt to interrelate and shape the information in ways that might make invisible histories

much more visible, create models and visualizations about historical questions, and attempt to uncover patterns and relationships not otherwise apparent.[11]

What follows seeks to place our Digging into Data project into context and to explore the social and methodological questions we confronted. We take the view that our project opened up key questions in Digital Humanities about the challenges and possibilities of doing "big history"--the use of large-scale data in historical inquiry. The challenge before us is significant: Google Books has digitized over 15 million volumes, the Clinton White House generated over 40 million emails, the Library of Congress' digital newspapers project has digitized hundreds of millions of page images. Even relatively small data sets can contain big sets of associations within them. Bringing this issue into focus for our project, for example, we found over 8,300 geographic place names used in just four years of one Civil War newspaper (The Richmond Daily Dispatch). But these unique place names were used over 292,000 separate times in the newspaper during the war. Working with extremely large data reveals that we will need to adopt different approaches and research methodologies, as well as structural changes in our practices of scholarly communication. Most of all, humanities scholars will need to develop tools that contextualize otherwise abstract data points, to reveal how they are embedded in local and specific contexts.[12]

As much as the Google Books project opens up the possibilities for working with large data sets, the records of a social process and technology as complex as the railroads remain difficult to access. The recently issued Council on Library Information Research (CLIR) report on "Mass Digitization" cautioned that monographic literature represents only one type of text scholars analyze: "Humanities scholars will increasingly want to do much more with text than use it simply as an alternate format to print. They will want to mine and recombine it, which is not possible with the current products of mass-digitization projects. Indeed, future reading will be done in part by machines in such a vast repository of information." Robert Darnton in calling recently for a National Digital Library project is responding to these concerns. Clearly, scholars need access to "big data" texts for

research, but we are a long way from having reliable, open-source, rights-free, and richly encoded texts.[13]

Multi-Dimensional Inquiry

The present Railroads project, perhaps rather more than most of the other topics funded in the first round of the Digging into Data Challenge, requires not only the analysis of very large datasets individually, but also the *integration* of multiple datasets of very different types. Specifically, these include unstructured text from both scanned books and newspapers, structured alphanumeric datasets from census and non-census sources, existing GIS datasets, and large numbers of both digital and paper maps. The key point is that the project is first and foremost driven not by technology (though it uses a great deal of it) but by the desire to address substantive research questions related to the nineteenth-century railroad industry that conventional scholarship from the pre-Digging era has been unable to address adequately. It therefore attempts to provide some case studies of effective utilization of new technology, large data volumes and subject expertise, in a manner that is accessible to a broad range of researchers whose interests are primarily substantive in orientation. All aspects of data assembly, integration and analysis, are therefore deemed appropriate topics for investigation, with a view to communicating the lessons learned to a wider audience.

To this end, it is useful to conceptualize the multiple dimensions that characterize each type of on-line research resource, since the meshing of these dimensions for different resources will determine the 'feasible region' where data integration is potentially a fruitful activity. These possible dimensions can be grouped under several headings, as follows, though they will vary in importance, depending on the type of resource in question:

- Availability and accessibility
- Quality and comprehensiveness (temporal and spatial coverage)

- Coding, tagging and consistency/standardization
- Metadata and documentation
- Existence of suitable 'authority files'

Taken together with the substantive research questions of interest, it should be possible to assess the usability of individual resources, based on these criteria. However, for the integration of multiple resources, there must be sufficient 'common ground' between the relevant dimensions for each resource, to make appropriate kinds of data matching, linking or correlation (in the broadest sense) possible. In the simplest and most obvious of cases, if there is very limited overlap in either time period or geographical coverage between two datasets, then the process of integration may yield very few useful insights. In other more complex cases, significant preparatory work may be necessary before the potential benefits (or otherwise) of possible integration can be determined.

With a mix of structured, unstructured and image datasets, the resulting integrative possibilities must then be subject to further evaluation in the context of available technological options for processing, analysis and visualization. Once again, certain technologies may be suitable for specific types of processing on individual resources, but be of limited value for data integration, while for other software tools the reverse may apply. There will also be gradations in the relative importance of manual versus software driven approaches, remembering that the driving force is improved answers to substantive questions, and the harnessing of technology to that end, rather than the pursuit of technology for its own sake. This may result in perhaps unexpected but effective use of human-machine combinations, targeted specifically at adding research value to existing on-line resources.

Availability and Accessibility of On-line Resources

One of the notable aspects of 19th century imprints pertaining to the US Railroad industry is the

very large number of different annual series of published works. These range from standard annual reports of individual railroad corporations to the voluminous tomes produced by State Railroad Commissions. Many of the latter and proportionately much fewer of the former are now available from a combination of Google books, the Internet Archive and the Hathi Trust. However, nearly all the series are incomplete to greater or lesser degrees, some volumes can only be read on-line and for the vast majority of Universities which are not yet subscribers to the Hathi Trust, readers can only download 1-page pdf files, just about acceptable for selective printing but not for any kind of automated use. Very significant time and effort is therefore still required to locate, and where appropriate download relevant pdf files for subsequent use. Other impediments to efficient working include the weak bibliographical referencing on the Internet Archive which makes identification of specific volume holdings difficult, and the tendency for collections of annual reports to be scanned as single volumes without reference to the fact that the items contained therein may not form a complete run and indeed may not even be in chronological order.

Given this patchy and often extremely limited coverage, especially of company annual reports, over the course of this and earlier projects, the University of Portsmouth has developed a large digital, microfilm and indeed paper library of railroad materials and maps, especially focused on rare and early imprints found in libraries such as the Hagley Library, the Historical Society of Pennsylvania, the Baker Library at Harvard Business School, the John W. Barriger III National Railroad Library at the University of Missouri, St Louis and the Northwestern University Transportation Library. In terms of this project, these resources are complemented by the very substantial Kennedy Railroad Collection at the University of Nebraska, Lincoln.

The position with scanned railroad maps has improved considerably over the last five years, led by the Library of Congress and the David Rumsey collection, though the University of Alabama and other non-academic sites are now making important contributions. Interestingly, these on-line collections contain surprisingly few of the important map series found in State Railroad Commission

Reports. It should be noted in this context that the Google project does not scan folded maps attached to books or reports, which is a particularly unwelcome omission for railroad research. In consequence, the University of Portsmouth has targeted over a period of years the acquisition of original State Railroad Commission maps for the Northeastern USA and its collection now rivals the British Library as the leading European holder of this particular type of source document.

Experience with scanned newspaper collections raises a number of questions in relation to accessibility and availability of underlying data. Many of the world's leading libraries are now involved in initiatives of this kind, such as the Library of Congress and the British Library, as are state and major university libraries. However, access to on-line collections has largely been designed with the needs of individual readers in mind, rather than the requirements of automated and exploratory data analysis tools. Further to this, the complexity of the underlying digital rights and proprietary aspects of indexing technologies, may preclude the possibility of gaining access to large bodies of scanned newspaper text. This has proved to be the case for the present project in relation to the Pennsylvania Civil War Newspapers collection, for example, and it is clearly a matter in need of attention by funding bodies for digitization projects of this type.

Structured tabular data, whether derived from census or non-census documents, is another important type of on-line resource. The main repository for census data has clearly become the Minnesota Population Center (MPC). The US National Historical GIS (NHGIS) and the North Atlantic Population Project (NAPP), both run by MPC, are of particular significance in this context. The former provides county level tabulations from the printed census volumes, the latter the 100% count of the US 1880 census, together with sample counts from earlier and later censuses and a series of samples recently linked between pairs of census years. Some academic and genealogical websites may also contain transcriptions of manuscript census schedules for specific counties, though coverage is very patchy both over space and time, and permission must be sought for any downloading and storage in databases. Ancestry.com contains an immense wealth of original manuscript census schedules,

accessible on a subscription basis, but the user agreement precludes large-scale systematic data collection from these primary source materials.

For non-census materials, be they in printed or manuscript form, the position is far less satisfactory at present. Small numbers of city directories for a handful of cities are now available from the Internet Archive or academic sites, with the notable exceptions of Baltimore and Pittsburgh, where 19th century coverage is good. However, these are provided simply as scanned texts rather than structured datafiles, though in the case of Pittsburgh the resulting OCR has been indexed. Nor are county genealogical websites much assistance in this regard – only Ancestry has a very useful collection of city directories, especially for the period surrounding the missing 1890 census.

The situation for data derived from manuscript materials, e.g. company records or payrolls, is less satisfactory still. Only tiny amounts of material are extant on the web, e.g. employee lists of railroads or other industrial concerns in specific locations. The present PIs have been at least as active as any other individuals in trying to address this situation, as will be seen below, but the amounts of available data are still very small. Only in the case of Pennsylvania coalmine accidents from 1869 onwards has a significant corpus of on-line material been built up (> 80,000 records) and one of the present PIs has contributed to this overall effort as part of a previous ESRC-funded project (see the on-line resource accessible via www.nehgis.org). However, attempts to secure access permissions to download additional data from genealogical providers have not proved successful to date.

The notion, therefore, that vast corpora of on-line data are readily available for Digging-type projects is a substantial over-statement of the present situation, especially when the requirements of specific research investigations are taken into account.

Quality and Comprehensiveness

The additional dimensions of quality and comprehensiveness place further constraints on the

potential scale and scope of Digging projects. Simple examination of the text versions of scanned books shows that the quality of the OCR is distinctly variable, often as a result of the irregular and broken nature of many 19th century fonts and the poor overall print quality of many 19th century and earlier works. For broad-brush studies based predominantly on word frequency data, this may not be too severe a problem. However, for studies that require accurate matching of names, for example, even a limited number of erroneous mismatches may hinder the investigation. This much is quite well-known, though systematic assessments of OCR error levels across different types of material are hard to find. Less well-known is the fact that most OCR programs perform comparatively very badly on complex tabular data, especially when original fonts are small, and the resulting output is often not only unusable, but also unintelligible. For railroad and other company annual reports, where half or more of the printed pages may contain tabular material, this is a severe, if not insurmountable impediment to any kind of reliable automated analysis of the numerical data they contain, at least with present levels of OCR technology.

Similar quality concerns can arise with newsprint, which is often scanned from microfilm, thereby introducing a further source of quality loss, or because tight bindings of original volumes prevent satisfactory scanning of left- or right-most column inches. In the case of city directories, the multi-column format has caused problems for early OCR work, leading to information from different entries on different sides of the page being concatenated, but this seems to have been alleviated in, for example, the more recent text files for the Baltimore city directories. Such differences can have an important impact on how the resulting digitised files can be processed.

The quality of digitisation of US census datasets, as opposed to the inherent quality of the data themselves, has not yet been the subject of extensive investigation, in part because many of the datasets are of relatively recent origin. Sources of error that have come to the PIs attention when checking against original schedules include transcription errors, misunderstanding of occupational descriptions and failure to take account of important marginalia, e.g. that the persons listed are penitentiary

occupants. The level of transcription errors for surnames has been found to be quite low in data derived from Ancestry.com sources (such as the US 1880 census) when the quality of the original schedules is reasonable, but the errors when made would usually be sufficiently great that they would result in actual matches to names from other sources being missed, so their impact is not negligible. It should be stressed, however, that these findings are based on impressions gained in the course of other work on this project, rather than as the result of a systematic process of sample checking.

Comprehensiveness has been touched on in the previous section and it is already clear that there are major gaps in the current levels of digitization of railroad reports, often mediated by the relative extent of holdings for individual companies in the Pliny Fisk library at Princeton. The latter institution has been a major and valued contributor in railroad terms both to the Hathi Trust digital holdings and those on Google Books. Unsurprisingly, larger companies with greater longevity are better represented than the smaller and more ephemeral and there is better coverage for the 1880's and 1890's than in the ante-bellum period. Overall at present just over 500 entries are returned by a search for 'railroad annual report' in the Hathi trust catalog, though a number of railroads have multiple entries owing to minor name changes after re-organizations. This implies about 6% of the railroad companies ever chartered in the Unite States are currently represented, enough for significant sample work on textual data to be undertaken but unsatisfactory for studies based on specific geographical regions, especially in the mid 19th century.

The comprehensiveness of the 100% count 1880 census should not be an issue, though under-enumeration may be. Similarly, the MPC 1% and 5% samples for other censuses are designed to be as representative as possible. However, the degree to which linked samples between censuses are as representative is more open to question, as the success of linkage varied both between geographical regions and between occupational groupings, sometimes to a marked extent (Goeken 2008, pers. Comm.). Assessing the comprehensiveness of city directory data is very difficult because it is not always clear exactly when the data were collected, so even in a census year this could result in

significant differences in the content of name lists, as compared to the census, from this cause alone, especially in urban areas where the population in general may be more mobile. In non-census years, in the absence of detailed tax records, there may be no obvious independent means of checking the completeness of directory entries. Company payroll records, though hard to find, ought to be more reliable, since payments were made on the basis of the names listed, though they can be sometimes be complicated by missing/illegible entries or payments made for work done in previous months.

Overall, a reasonable range of scanned maps for most areas is now available online for a range of dates. Some areas, such as Kentucky or Virginia in the 1880's-1890's, have rather limited coverage, however. There has also been very little work undertaken on the accuracy of 19th century railroad mapping. Problems of accuracy can take many forms, including showing lines as completed when they are still under construction, incorrect drawing of routes between specific locations, omission of depots for ease of labeling, and partial updates, so changes shown do not correspond to actual changes in the network by the time of the copyright date. A by-product of the current project, which produces large-scale GIS outputs, but always based on multiple sources, some of them non-cartographic, is that detailed checks on the accuracy of historical cartography at specific dates can now be made.

Coding, Tagging, and Standardization

The coding of census or city directory data to standardize occupational characteristics or industrial affiliations is much more fraught with difficulty than might at first appear. The 1880 census employs a modified version of the HISCO historical occupations codes, but this has been found both in this project and a previous one funded by ESRC, to be inadequate for the proper description of US heavy industrial occupations in the 19th century. For example, fewer than 10 codes are used for the coal mining industry and only 22 for all railroad related occupations when payroll and city directory

analysis of workers in these industries has identified hundreds of different occupations working in these sectors. An even more serious problem relates to the assignment of individuals in different occupations to specific industrial sectors. Although the NAPP dataset contains these industrial codes, it has been discovered that the occupational descriptions in the original census schedules do not in fact contain enough information to allow this assignment to be undertaken in many cases. This can be traced back to inadequate guidance on the collection of occupational data given in the enumerators' instructions. It has now been shown that many of the industrial codes assigned to workers in the railroad and mining sectors, for example, are erroneous for this reason. It has also been found that the problem relates to workers in generic trades, such as machinists, blacksmiths and indeed day laborers, rather than to industry specific workers such as coal miners and train conductors. Very large numbers of generic workers only record their generic trade in the census, not their industrial affiliation, so, for example, the overwhelming majority of machinists working for the railroads in the northeastern USA are not coded by NAPP as working in the railroad sector. Further details of these problems are described in a forthcoming paper in Historical Methods, but it should be noted they have a major impact on the types of analysis that can actually be undertaken on industrial workers in the census, without risking serious error in the resultant findings. The only long-term solution is to recode census entries using matched data from non-census sources, which contains adequate information on industrial affiliation of workers at the time of the census. This is a very difficult undertaking, although some methodological progress in the right direction is already being made.

A quite different set of tagging problems face investigators attempting to work with historical texts in general and newspapers in particular. Many of these revolve around correct attribution of proper names, when the meaning is heavily context dependent. Examples would be personal names that are the same as the names of towns, multiple towns or counties with the same name, and distinguishing between references to places that imply movement or journeying from other references to the same places that do not. Other cases of difficulty include reference to companies using

shorthand names, e.g. the ‘Baltimore Road’, which might or might not refer to the Baltimore and Ohio Railroad depending on the context. While these entries might constitute a minor, but perhaps expectable annoyance to researchers undertaking keyword searches in a traditional manner, they can introduce a large amount of 'noise' into a automated visualization system that is attempting to display large numbers of such references in a meaningful manner.

The clear lessons to be drawn from even this very brief overview of major coding and tagging issues is that scanning and OCR is simply the first stage in preparing large datasets for on-line use. Multiple further stages of data cleaning, standardization, and coding/tagging are likely to be required, before anything beyond basic keyword searching or word frequency analysis can be undertaken. And the greater the degree of required integration with other datasets the closer the attention that must be paid to ensuring consistency in coding and tagging activities.

Metadata and Documentation

Some of the problems of inadequate bibliographical metadata have already been alluded to in relation to the identification of actual annual coverage in scanned railroad annual report files. Most available scanned maps are available from authoritative sources, in cataloguing terms, such as the Library of Congress and major University libraries. However, the search mechanisms, like in the digital newspapers case, are oriented towards the needs of individual researchers rather than automated search or metadata harvesting methods. While catalogue fields will generally provide information on the map compiler and a corporate name, if the map focuses on a specific railroad line, other information on the actual map content and provenance may be lacking. For example, railroad maps vary considerably in whether depots or lines under construction are shown, and it may be important to know if the map was originally published as part of an annual report, an investment prospectus or indeed the report of a stockholders’ investigating committee. In consequence, time consuming visual

inspection of most digital maps is still required to determine their possible usefulness for specific kinds of work in conjunction with other sources. The growing importance of accurate metadata in support of federated search of multiple repositories for digital materials can be seen from the recently launched 'Connected Histories' website in the UK (http://www.connectedhistories.org/).

In the case of structured datasets, determining the availability of digital census data for different years is unproblematic, thanks to the efforts of MPC and other international providers of equivalent material. However, within datasets, good documentation becomes essential for effective use and again current systems are generally oriented towards manual selection and subsetting of data, followed by batch custom data file creation for later download. Using the current NHGIS interface, for example, it can be difficult to identify where in the search tree particular variables can be found, not all areas of census coverage are included (e.g mining industries), and, for checking purposes, it can be extremely difficult to determine exactly where specific data items in the digital corpus can be found in the printed census tables of the original published census volumes.

A different class of problem arises with the individual level data from the 1880 complete count dataset. The substantial number of coded original and derived census variables (80+) for each individual record means that there are a substantial number of code tables on the NAPP website, whose printed contents extend to a pile of paper one inch thick. These have been set up for selective inclusion with downloaded data files designed to be input to statistical packages. However, database users who require a full set of code tables are not adequately catered for and extensive data manipulation is required to process these tables into the required structures, so they could be used as part of more automated procedures. Further to this, while issues with occupational/industrial coding have been discussed earlier, additional problems of error and anachronism have been identified in the code structures used for geocoding purposes. Once again, data providers will need to give increasing attention to the quality, standardisation and on-line accessibility of coding structures, as these provide the keys to unlocking the information content in large datasets and establishment of degrees of data

comparability with other non-census sources.

Authority Files

Traditionally the preserve of the librarian, authority files can be expected to assume an increasing importance as exploratory data analysis extends over wider and wider bodies of digital material and increasing attempts are made to develop automated methods of tagging or geo-tagging both structured and unstructured datasets. Place names and corporate names are two obvious examples, where there is benefit in having ready access to the correct spellings or correct forms of names for purposes of correction, standardization or disambiguation. Unfortunately, this can often be done more effectively on a manual basis than in an automated fashion, because of the importance of context, that has already been stressed in an earlier section.

A classic example of the need for an authority file is provided by the highly complex chronology of railroad corporations in the USA during the 19th century, where corporate names frequently changed, either because of financial re-organization or company takeover. Tracing the lineage of predecessor, successor and subsidiary lines is a monumental undertaking, owing to the thousands of companies involved and the extent of changes in ownership or control (we have found one company with nineteen changes of this kind in the 19th century alone). Faced by such a daunting task and the amount of detailed research that may be required to document all such instances accurately, even the Library of Congress has faltered in this particular case. While many of their authority file entries are good and quite comprehensive (see 'Norfolk and Western Railroad'), others, such as the entry for the Cleveland, Painesville and Ashtabula Railroad, are incomplete, as they do not list successor railroads. Establishing correct charter names and their dates of applicability is difficult enough in its own right, but for purposes of linking different resources e.g .travel diaries and newspaper reports, lists of known name contractions, abbreviations, nicknames and wrongly-ordered names that

correspond to specific legally correct names are also required. The ease with which names can be wrongly ordered is apparent from examples such as the Cleveland, Columbus, Cincinnati and Indianapolis Railway. While assertions of completeness would be foolhardy under the circumstances, it is expected that current and future work by both partners in the current project will contribute to improvements in the availability of reliable authority lists for significant parts of the 19th century rail network.

Overall, this brief survey of the multiple information dimensions of different on-line resources only serves to highlight the crucial importance, not just of accuracy in the scanned documents and digitized datasets themselves, but also in the associated information superstructures that make them accessible to digital researchers and the kinds of software tools they will increasingly wish to employ.

Digging into Data Apps

Although the technical specifications of our approach to these "Apps" will be examined in detail below, we decided at an early stage to focus our development--technical and analytical--on the production of "Apps." These packages of data, interpretation, and metadata focus on using a type of data whether structured or unstructured, textual or tabular, in the service of a research problem or question related, in our case, to railroad development. The each app is composed of four elements: 1) an interaction interface that allows users to view the data spatially and temporally and relate it with other data in the system, 2) a data summary page that provides users access to raw data elements as well as summary information for how the data was curated for visualization, 3) a research results section that discusses how the data were used as the basis for scholarly interpretation of some historic event or concept, and 4) a developer section that provides documentation on how data can be accessed and consumed by other software projects. The goal of each app is to provide a 4-dimensional view of a data element and its relationships to other data, serving as a community portal for information discovery. A fifth component to each app is a data curation interface that allows members of our

research team to edit data with edits appearing in real-time in the app.

Our team concentrated on five Apps designed to enable "big data" inquiry framed around well-known historiographical questions.[14] These include:

- Employment mobility of railroad and other heavy industrial workers - a census-based structured data approach
- The Civil War and mobility - unstructured newspaper data extraction and visualization
- Network evolution - spatial data temporal representation and integration
- African American mobility after emancipation - approaches to data integration
- Landscape and environmental transformation with railroad development - text mining application using railroad annual reports

Railroad Employment and Mobility

This App addresses the first of the substantive areas of research focus identified in the original project proposal, namely railroad-related employment and mobility. Several more specific research questions were also identified, including the following :

How did the arrival of the railroad change the actual and perceived landscapes of employment opportunity, both on the network itself and in the communities it touched?

How did railroad-related labour mobility change in response to business cycle booms, strikes or other key historical events and were there differential effects by ethnic group or type of worker?

What was the temporal and spatial pattern of interchange of skilled and unskilled workers between the railroad sector and other heavy industries, since they shared a large group of common occupations, including blacksmiths, carpenters, engineers and machinists, who did not necessarily report their railroad affiliation in the census?

Aspects of these research questions are being investigated in the two regional case study areas identified, namely the Great Plains and the Northeastern USA.

Railroad employment became a leading engine of migration and movement within regions, between regions, and across national borders. In the broadest sense British capital, laborers (i.e. Irish), iron, rail, products, and technical know-how flowed across the Atlantic into the U.S. and U.S. goods, technologies, and securities flowed back into Great Britain. In addition, Mexican, Chinese, and

Canadian laborers came to the U.S. to construct the rail lines and migrants from all over Europe purchased railroad lands in the late nineteenth century and moved into the U.S. at the peak of transatlantic immigration. Yet, we know little about the spatial movements of the railroad workers, except in the broadest terms. More detailed analysis has been limited for two reasons. First, few railroad payrolls and employment records remain for the mid-nineteenth century, and even fewer contain information about previous employment, birthplace, or ethnicity. Second, the U. S. census data on occupation has been imprecise and often failed to capture whether workers in generic trades, such as carpentry, were employed by railroad companies. Third, race was not recorded in the aggregated 1883 U.S. Census reports and may not be recorded in payrolls, so this presents an especially difficult challenge for historians of African American history.[15]

In the original nineteenth-century reports, the U.S. census data on railroad employees was aggregated at several levels. First, in 1870 the U.S. census helpfully listed in aggregate form railroad workers for each state with data on ethnicity, but no county level information was provided (and race was not included). Then, in 1883 in its volume on special transportation statistics the census reported the number and occupational categories for each railroad company, but the tables were aggregated only by traffic groups (sub-regions). The most detailed railroad employee data from each company, therefore, cannot be mapped to U.S. states and counties. And in 1890 the census added race as a variable to the count of U.S. railroad workers for the first time. Curiously, in 1890 the U.S. census created a category for ethnicity from unspecified "other countries" and provided no data on China, Japan, or Italy, even though in 1870 the exact number of Chinese and Italian railroad workers was documented for each state.

Some railroad company records give more detailed accounts of the previous work histories for employees on the Great Plains. The Burlington's strike records described how many men came on as replacements in 1888 and where they came from to replace strikers. Hundreds of men were recruited from the East, but many were hired locally, and some came from California and places farther west.

The strike replacements possessed years of experience on over forty railroads. Several men worked on railroads in India, a few in London. The Burlington also began compiling a list of all dismissed workers on its lines. Between 1877 and 1892 over 8,000 employees were fired for causes ranging from carelessness to alcohol consumption. A further 2,000 employees were fired for strikes or labor agitation, many of them in the engineers and firemen strike of 1888. The spatial histories of these workers indicate both the dense concentration of experience in the North East and pockets of experience from other regions and locations, especially on the Great Plains.[16]

The Burlington was not the only railroad that began keeping such detailed records of employees and their work histories. The Union Pacific Railroad also collected detailed records and compiled lists of employees dismissed beginning in 1889. Over 1,000 employees were dismissed on the Union Pacific in 1889-1890, the largest number for carelessness or negligence (11 % of the total). An equally large number were dismissed for alcohol violations, as well as operating errors, financial mismanagement, and rules violations. Each of these causes for dismissal were spread across the Union Pacific's system, but striking was highly concentrated. From the Burlington workers' histories and from the Union Pacific's dismissal data, we can begin to see that railroad workers were highly mobile, and their work was characterized by spatial differences and spatial histories.

Our Digging into Data project has begun assembling a data framework for interrelating employee data from payrolls and censuses, attempting to reproduce the systems with GIS data on rail lines, depots, junctions, to make visible patterns otherwise invisible.

The most complete and detailed study of railroad worker mobility on the Great Plains has been Sheldon Stromquist's A Generation of Boomers. He looked at the changes in labor mobility patterns between 1877 and 1894 and its effect on railroad workers' social and geographical mobility. He argued that railroad workers movements were largely defined by an east-to-west pattern, but, during the Burlington strike, the character of this pattern transitioned from voluntary to non-voluntary migration. Stromquist argued that in the first phase of railroad growth--the "generation of boomers"--was

governed by the open market and the individual desire to move up in the industry. But this phase gradually drew to a close, because it was effaced by a regime of railroad company labor practices in the West, especially the personal card system, blacklisting, promotion from within, company welfare policies, and the hiring strike breakers from other places. This series of company practices altered the railroad work environment which both shaped and was shaped by the strikes of the period.[17]

Stromquist focused on how geographic and social mobility interlocked. He looked at two Iowa towns--Burlington and Crestone--and used a unique set of records to define railroad places. Iowa conducted a five-year census which collected detailed occupational information. Stromquist determined that there were twenty-four railroad towns in Iowa. He chose places only with more than 100 railroad employees based on this census, but these twenty-four towns contained 93 % of all railroad employees in the state. Stromquist then compared the individuals to their 1880 census data and was able, therefore, to determine ethnicity and age cohorts. He argued that Burlington was ten years younger than Crestone, and the age cohort of employees differed quite markedly. The result was a different response to the 1888 strike and the company practices in its aftermath. There are, however, few other states with such detailed censuses.[18]

On June 23, 1888 W. F. Merrill, the General Manager of the Hannibal & St. Joseph Railroad and the Kansas City, St. Joseph & Council Bluffs Railroad, reported to the Chicago, Burlington & Quincy's headquarters in Chicago on the aftermath of the strikes in April and May by engineers, firemen, brakemen and switchmen. The Burlington officers asked each manager for a detailed assessment, including who the strikers were, where replacements came from, which employees deserved special mention for their loyalty or service, what casualties resulted from the strike, and which newspapers and towns supported the strikers. There were eleven questions in all. Merrill reported that the superintendent of the Hannibal & St. Joseph, S. E. Crance, deserved to be singled out for his "devotion." Crance lived in Brookfield, Missouri, where "most of his engineers & firemen lived," a place that was "almost entirely a railroad town."[19]

Crance was applauded because "he was in a position where it was difficult to get hold of men to fill the places until they could be sent from the East." On the Hannibal & St. Joseph, Crance, like the other superintendents on the line, hired "new men" to replace the strikers. Men came from Kansas City, Atchison, Quincy, and other towns nearby. But they also came from Chicago, Grand Rapids, Cleveland, Philadelphia, and from California. On the Hannibal & St. Joseph 170 men participated in the strike and on the Kansas City, St. Joseph & Council Bluffs 115 men struck. Replacements were recruited quickly from local men who refused to strike. Three days into the strike these railroads were receiving "corps" of switchmen "from Philadelphia," ten at a time.[20]

In response to the great strike of 1888, the Burlington began to maintain a "personal record" on each employee, essentially a "blacklist." These records indicated not only whom the Burlington would not rehire in the wake of the strike but also the work history of the employees. Here are two entries:

"James Kain. Loco. Fireman
Commenced February 1883 as a Wiper.
March 1883, to Stack Inspector.
February 1884, to Machinist helper.
August 1877[1887], to Fireman.
Quit February 27th, 1888, account of Strike. Put in application on form 1608 January 31st, 1869.
Record during Strike. A slugger, got to drinking and set on by other men and was bad."

The Burlington determined that 17 % of the applicants would be classified and marked as "objectionable." Unsurprisingly, a red mark was placed near James Kain's record, and he was blacklisted.

"L. W. Giles, Loco. Fireman

Commenced February 1882, as Wiper.

April 1882, quit.

July 1885, returned as Clinker.

February 1886, Laid off, no work.

March 1886, returned as Wiper.

January 1887, to Boiler washer helper.

August 1887, to Fireman.

Quit February 27th 1888, account of Strike. Put in application on form 1608 January 12th, 1889. Record during strike. Very quiet during strike"

Giles was not blacklisted. No red mark appeared next to his name, but none of the participants in the strike were ever rehired by the Burlington--the reapplication process was largely a formality to appease the worker's committee that ended the strike.[21] A handful of engineers did not even bother to put in an application to be reinstated on the Burlington. They moved on to other places: one was "running on some road out of Denver," another was in North Carolina, one each went to the Northern Pacific and the Atchison, Topeka & Santa Fe, and one was "running a butcher shop in Burlington."

A great deal of evidence about railroad workers is contained in these short biographies. Men moved in and out of railroad employment. They quit, they were dismissed, suspended, fired, rehired, quit, rehired again, and struck. They started in one position on the payroll and then moved to another. Sometimes they moved up and then back down. Railroad work was highly specific and diverse, and some of the occupational categories were flexible. A brakeman could quickly become a fireman if given the opportunity.

One of the difficulties in assessing the railroad workers and their spatial mobility has been simply identifying how many workers there were and where they lived. What exactly was a "railroad town" or railroad place? Indeed, the 1883 special report of the U.S. Census showed 418,957 railroad

workers in the United States, based data from the railroad companies. But the total number of railroad workers who identified their occupation as railroad-related in the 1880 U.S. census was nearly 267,000. The occupational coding for the census has proven especially accurate, not surprisingly, for the conductors, engineers, and train men (The North Atlantic Population Project data uses the HISCO 506 industrial code for railroad and railway employees.) But for shop men, such as machinists, tinners, and boilermakers, and probably for carpenters, blacksmiths, and helpers, the coding breaks down, and individuals in these generic trades fail to show up as having railroad occupations in the 1880 census. For machinists a sample of 5,800 workers from the NAPP data produced just 6 % who were properly identified as railroad workers, even though railroad companies hired thousands of machinists.

To address this problem, we have created a railroad occupational index to account for the shop men as railroad employees and examine the locations of railroad-centers across the U.S. (see Figure 1 Appendix) The shop index includes machinists, boilermakers, tinners, and moulders--the most common railroad employees. We combine these individuals with the railroad men and divide the total number by the total number of males in each county in the U.S. We have identified all counties with more than 1,000 men in this count and more than 2.5 % of all males in the county as "railroad centers." The combined criteria prevents counties with very small populations and a few railroad men from distorting the results. By this measure 38 counties can be determined as highly concentrated railroad centers or places. The index proposed tends to favor locations with very high ratios of railroad men to total males, where they constitute a large proportion of a relatively small population.

Nevertheless, the national view of railroad centers drawn from the data (and otherwise impossible to observe) indicates how quickly and substantially the railroads concentrated in parts of the Great Plains. The emergence of Omaha, Nebraska, and Council Bluffs, Iowa, as major railroad centers, along with Denver, Colorado, was part of a larger pattern of Western concentration and intensity, even as the overall weight of the railroad occupational structure remained located in the North East. By this measure five of the eleven most concentrated centers of railroad employment in 1880 were located in

the Great Plains or West.

Railroad employment in the West differed from the East in several important ways. Railroads in the West were less densely interwoven and were separated by vast spaces. They could exert more control over where employment was offered and to whom. They possessed an unusually high degree of what we might call "spatial monopoly power." In the East by contrast, large numbers of small and large railroad companies competed for labor, especially among the shop trades and generic occupations. For the Great Plains, when we lower the railroad index to counties with greater than 500 railroad and shop employees, the density of railroad centers in the East becomes even more visible, and, in addition, the cluster of railroad centers along the Missouri River Valley in Nebraska, Kansas, Missouri and the Midwest.

Our national railroad-center index drawn from the 1880 NAPP data provides the capability to examine the national patterns in a way that earlier scholars could not accomplish, and therefore to place individual case studies and communities in context. Ethnicity, age cohorts, and race can be drawn from these data as well. Other opportunities will follow--to cross correlate among railroad places to see spatial histories of workers and communities unfold. We have already found new patterns in the area of African American employment on the railroads, data that would have been otherwise possible to obtain--that 52 per cent of all railroad workers in Virginia in the 1880 census were African American while 98 per cent in Maryland and Pennsylvania were white. We hope to develop tools for accessing and visualizing these otherwise hidden data.

The second of the case study areas covers a number of north-eastern industrial states outside New England where a substantial proportion of the nation’s railroad employment was concentrated, together with that in the mining and iron and steel sectors. Major sources for the work include the 100% count NAPP 1880 census records, other census records where available, city directories and company payroll records. Comprehensive data (including names) on males employed in five north-eastern states (MD, OH, PA, VA, WV) is already stored in the 1880 census warehouse developed on an

earlier ESRC-funded project (see Historical Methods paper).[22] Limited datasets from Baltimore city directories in 1858 and 1860 were also available when the project commenced, together with railroad payrolls for the Baltimore and Ohio, largely dating to the 1850's, and a large new dataset of the names of employees of the coal mining department of the Delaware, Lackawanna and Western Railroad (DLWRR) during the Civil War has also become available. The latter derives from a separate project funded by the Pennsylvania Historical and Museum Commission and one of the present PIs is technical advisor to the project. (The DLWRR was a leading carrier of anthracite from the coalfields of eastern Pennsylvania to Upper New York State and the cities of the eastern Seaboard.)

Mobility in the Civil War

The Richmond Daily Dispatch, like many newspapers during the Civil War, carried short notations of movements--prisoners were brought down the line from point A to point B yesterday, for example; or the soldiers of the 54th Virginia were seen moving toward Gordonsville from the depot at Bristoe. Refugees received prominent mention, and contrabands appeared in the Confederate newspapers. We initially sought to extract from the textual data any mentions of "from . . . to" events, and then to categorize them by type based on the keywords contained in the text and the associated place names and dates. Refugees, soldiers, deserters, contrabands, and women would be types. We also sought to collect the date of the event, and its reference. The visualization tool would allow a user to look for all movements of a given type in a given year, month, or day and to see these plotted on the framework map. Other data, especially railroad node and depot data, would allow a detailed and accurate graphing and modeling of human movements in the war.

As previously mentioned, we found in four years of the newspaper over 8,300 unique place names and these occurred over 292,000 times. Many of these needed correction. As of this writing, 3,085 places have been corrected and checked. Some names that referred to individuals, such as "Washington," were incorrectly identified as places. Then, other places which no longer exist, such as

towns in Tennessee since 1935 under water in the Tennessee Valley Authority, were coded against current geographic gazetteers and matched to places with the same name but elsewhere, for example in Ohio or Virginia. The largest set of errors were simply inconsistencies--Wise, Virginia, was coded as Wise County, South Carolina, in some instances but not others. Finally, non-U.S. places, such as "Ghent" or "Cerro Gordo" were wrongly identified as North Carolina not Belgium or Mexico respectively. These places needed to be laboriously corrected. Even using the cleanest newspaper texts, therefore, a straightforward problem in geocoding required significant data preparation, proofing, and correcting. Indeed, to make this possible, we created a data editor for quickly examining, checking, and geocoding place names against the text. The editor is now part of the general framework tool set, available for use and modification by any group seeking to check and geocode place names in xml-encoded texts.

A similar problem affects railroad company names. Over 370 individual railroads appeared in the four years of the Richmond newspaper, but each appears multiple times. Newspapers routinely shortened the names of individual roads. We have standardized these names to match our GIS of the 1861 railroad network, but some inconsistencies remain. Proper authority files for railroad names remains an area of research for this project.

The intense and accurate geocoding of the newspaper text allows us to create visualizations of keywords. Because we the xml encodes each sentence, we can determine where any place is used in a sentence with a given word. Contraband, slave, fugitive, deserter, and guerrilla, therefore, can be queried for the place names associated. We can represent the 781 instances of "contraband" of which 421 occurrences include a place name, including 157 unique place names. The word "slave" occurs with 570 unique place names. The App maps the geographic occurrences of any word, organization, or term.

Network Evolution

The Paullin Atlas map from 1932 has become the classic cartographic representation of how the railroads compressed time and space by 1857 to facilitate access to the continental interior.[23] The perspective centers on New York and implies a more homogeneous 'surface of accessibility' than was actually the case until at least the 1880's. In fact there were major accessibility variations across even those north-eastern states where railroad development was most advanced, including, for example, significant 'holes' in the surface in Northwestern Pennsylvania, western Virginia, and across wide swathes of southern Ohio. Likewise, the actual time-cost of overcoming distance was heavily mediated by line network connectivity (including gauge change problems), proximity to a changing landscape of depot construction, train frequencies, presence of single or double-tracked lines (the latter dramatically reducing timetable delays) and the availability or otherwise of express passenger/freight services.

In a sense, the combined effects of many of these factors are summarized (or even concealed) within the detailed tabulations of the large system timetables published by many of the major railroads. However, their geographical expression is very difficult to visualise without quite complex prior data processing of the timetables themselves and access to a detailed GIS of the rail line system to which they refer. Added complexity is provided by changing depots names, addition of depots over time along lines, and the impact of branch route sub-timetables.

This App aims to show how visualisation capabilities can be integrated with a time table database and the required GIS data, chronologically matched to show the requisite depot name and location information for the time period in question. Attention will also be given to the possibility of automating the capture of time table information into database tables. The Baltimore and Ohio network as far as the western boundary of Ohio will be used as a case study example, and accessibility analyses for locations other than New York, such as Baltimore and Columbus, will be undertaken to provide a comparison with the original manually generated maps from Paullin's early work. Accessibility to the rail network grew dramatically over the period but some places became less well-connected that they were previously.

A second aspect of the App is to show the changing "geography of control" across parts of the railroad network using the examples of the Baltimore and Ohio, the Erie, and the Pennsylvania Railroads. The detailed chronology of construction, takeover and ownership of individual short lines by these major trunk lines has been traced and stored in an Oracle database, using a combination of on-line resources and materials gathered from the Barringer Railroad Library. The chronological information held in the database and the geography of the network held in ARCGIS have then been matched up throughout the nineteenth century to allow GIS shapefiles of the evolution of these systems across the chosen states to be created so they can be visualized using the Aurora Engine. The accompanying illustration of the Baltimore and Ohio RR in 1900 (see Figure 2 Appendix) shows the complex range of control strategies employed by the railroad to expand its network in different parts of its region. The apparent gap on the map in Virginia is not an error -- the railroad lost control of this portion of the line between Strasburg and Harrisonburg in the course of its financial problems and subsequent re-organization during the 1890s.

African American Mobility after Emancipation

One of the most important economic arguments in Southern history concerns the relative lack of mobility for African Americans during Reconstruction and the New South periods. In this view black labor was trapped in the South after the Civil War, prevented from northward migration and therefore open competition by northern racism. Other factors may have played a more important role, however--such as the changing composition of the railroad network across space, the birthplaces and family connections of mobile populations, and the building and construction of railroads in the region. Only World War I and the Great Migration changed this pattern of restricted geographic mobility for blacks. Yet, we know that considerable intra-regional migration took place, as freedmen moved across the South in the aftermath of slavery and by the 1880s were moving with the railroads into new opportunities made possible by the railroads.[24]

In fact, considerable intra-regional migration did take place, as freedmen moved across the South in the aftermath of slavery and by the 1880s they were moving into new locations made accessible by railroad developments. This App seeks to combine county level data on population changes and racial composition with detailed railroad network growth data and the records of the Freedmen's Bureau in key railroad centers.

Railroad linkages between North and South were limited to four or five gateways before the Civil War--places where the South's railroads linked to roads into the Northern states: Alexandria, Virginia, Cairo, Illinois, Cincinnati, Ohio, and Louisville, Kentucky. We sought to combine county level population changes (race) with detailed railroad network growth data to understand the shape of mobility after emancipation. When we examine the Freedmen's Bureau labor contracts for gateway cities--Alexandria, Virginia, Petersburg, Virginia, Louisville, Kentucky, Chattanooga, Tennessee, and Memphis, Tennessee--we see different patterns of work placement over time and across space.[25]

We have entered all Freedmen's Bureau labor contracts from the following railroad places:

Alexandria:	594 records, September 1865 through end of December 1866
Camp Nelson:	105 records, May 1865 through November 1865
Chattanooga:	426 records, February 1866 through April 1866
Louisville:	124 records, June 1866 through March 1867
Petersburg:	128 records, July 1865 through November 1865
	1381 records total.

We have no firm conclusions at this stage about mobility, but instead present here different tools for assessing where to begin looking in greater detail. It should be noted that nearly all long-distance labor contracts included a transportation voucher for travel by rail. In addition, some large numbers of contracts for dozens of freedmen were given to individual railroad companies, such as the Memphis & Charleston Railroad Company.

In addition, the startling difference in railroad employment in Northern border states and

Southern states revealed in our analysis of railroad workers in the 1880 NAPP data explains some of the potential limits to mobility after the Civil War. The racial breakdown of industrial workers by county and state represents an entirely new finding, otherwise unreported in the U.S. Census aggregate reports.

This App allows us to isolate one aspect of the process of mobility after emancipation--the labor contracting activities of the Freedmen's Bureau and visualize their reach, scale, and sequence over time. In addition, we expect to layer other relevant data on population change and railroad growth to better understand spatial patterns and correlations. In the absence of location-specific data about the movements of African Americans in the years immediately following emancipation, historians have relied on textual accounts. This App seeks to hold a framework for further additional data.

Landscape and Environmental Change

The building of railroads prompted large-scale landscape transformation. Wharves, bridges, rail yards, quarrying, timber harvesting, coal consumption, and tunneling proved especially important. Ties and timbering for bridges were brought from other regions and by the end of the nineteenth century railroads were replacing millions of sleepers(ties) each year. We have little overall sense of how resources were extracted in the wake of and in relationship to railroad construction. On the Great Plains the building of the Union Pacific and later the Burlington system opened up coal mines in Wyoming, rock quarries in Utah, and timber harvesting across the Rocky Mountain West. Each railroad possessed huge grants of land, timber, minerals, and water. Yet, historians have few comparative frameworks for documenting and analyzing these changes.[26]

This App focuses on the mining of textual sources, especially railroad annual reports, and secondarily newspapers, to document and index landscape change on a given railroad system. Some runs of railroad annual reports are now available on-line from the Internet Archive (Internet Archive 2011) and we have identified a suitable run for one of the Western roads. Such reports routinely

tabulate and narrate the use of timber and other resources in the construction and maintenance of the railroad.

The newspapers from the Nebraska Digital Newspapers project were evaluated for runs in the 1880s, but the results proved disappointing. Our project partner in Computer Science at Northwestern University, Assistant Professor Doug Downey, examined a set of sample METS/ALTO xml files and found twenty to thirty per cent of the words were incorrect. More importantly, the names and named entities, including places, were particularly inconsistent and incorrect. A sample of the "dirty" OCR included the following sentence: "'So It would seem,' chimed In Petty, 'but if ho mid been with me'n the late General Gcoi'KO Crook up at Horscshoo luke that Marcli afternoon in S3 when we had our llttlo mntlnoo with the canvnsbacks...'" Given the time constraints we faced on the project, we concentrated on railroad annual reports and set aside the error-ridden newspapers.

We have downloaded the following reports from Internet Archive for encoding place names, organization names, and dates related to construction, resource events, and resource extraction:

Central Pacific 1878, 1882-1886,	307 pp.
Virginia Central Railroad Company 1861	20 pp.
Gulf, Colorado, and Santa Fe Railroad 1883	187 pp.
Marietta and Cincinnati Railroad Company 1851	52 pp.
Pennsylvania Railroad Company, 1853	86 pp.
Pennsylvania Railroad Company,1872-1877	929 pp.
Chicago, Burlington & Quincy Railroad Company 1855-1870	443 pp.
Norfolk and Western Railroad Company 1890 - 1895	377 pp.
New York and Erie Railroad Company 1853	133 pp.
Pacific Railroad (Missouri) Publisher: St. Louis, Mo. 1851	144 pp.
Boston and Worcester Railroad Corporation 1844-1852	274 pp.
Central Southern Railroad Company, to the stockholders, 1861	20 pp.
Delaware, Lackawanna and Western Railroad Company 1854-60	387 pp.
Union Pacific 1880 -1883	275 pp.
Total:	3,634 pp.

Each of these reports is being marked up with place name, date, and organization name encoding for:

Type: Resource Use
Name: Coal, Oil, Steel, Timber, Stone

Type: Resource Extraction
Name: Coal, Oil, Steel, Timber, Stone

Type: Resource Event
Name: Fire, Material Flow, Flood

Type: Construction
Name: Activity, Bridging, Tunneling, Grading, Tracklaying

Aurora Engine Technical Development

In order to accomplish the evaluative goals of our answer to the Digging into Data Challenge, we have applied software engineering and computer science best practices in data integration. The result of this effort has been the development of what we are calling the Aurora Engine. Based on previous introductory work as part of the Aurora Project, the development of the Aurora Engine extends our initial spatio-temporal visualization framework by introducing a so-called data bus, allowing large quantities of disparate data to be quickly integrated and commonly visualized and manipulated across space and time. The introduction of manipulation alongside visualization has supported much of the historical scholarly research conducted during this project, and will serve as the basis for future developments. While not live at the time of this writing, a newly designed Aurora Project, powered by the Aurora Engine, will be a freely available artifact of this project and will provide both visualizations, scholarly interpretation of data, raw data, and Application Programmer Interfaces (APIs) to allow other projects to build off our ongoing efforts.

The Aurora Project takes a unified approach to the presentation and manipulation of data elements provided by the Aurora Engine. Each core data component – discussed in the sections above – is modeled as a internal web application component, or an 'app.' Figure 3 (see Appendix) presents an example of the interaction interface for Richmond Daily Dispatch Newspaper app. The primary

interaction mechanisms are a map depicting references to places found in an issue of the Daily Dispatch for a date. The user can adjust time using the calendar or timeline widget, and adjust spatiality using the map. Occurrence data is displayed spatially, on the map, and temporally, on the chart. The presentation of results is uniform across all of the apps, allowing for a single user interaction model and providing consistency to support information discovery and analysis.

Supporting our end user application model is the Aurora Engine, a distributed computing framework that supports data integration and access. The framework serves as a data bus, connecting data of differing formats from disparate sources to allow real-time access to integrated artifacts from the raw data. This approach makes possible the dynamic web visualizations and analysis techniques that are available in our web interface and differ dramatically from conventional data gathering and presentation efforts in many digital humanities projects. Real-time access to raw data components allow visualizations to be dynamically built based on user inputs – in our system those are time and space – and related to other data that are available in the system. This focus introduces a data-mining component to web application tools, providing a high-level of sophistication with minimal end-user complexity. The underlying data bus has the additional benefit of abstracting the actual tools from their supporting data, which allows for future data sets to be easily incorporated without needing to change the higher-level app. For example, as rail network geography are digitized for "new" rail lines in the future, their inclusion in the applications is a matter of updating the Aurora Engine dataset; no additional application development is required. The abstraction of data from tools will support considerable extension in the future and allows the project to easily incorporate feedback from advisors and partners.

The Aurora Engine is organized as a cloud-computing framework and employs standard techniques for distributed computing. While this fact is generally not readily apparent to the end user of the system, its properties have significant impact to the way in which the user is able to interact with the system. In a cloud-computing environment data and services are made transparently available to

the user through the internet (the "Cloud"). These data and services may come from multiple sources and be consumed on multiple devices at any given time. The key concept is one of transparency; the data are available when they are requested and the infrastructure to support availability is black box. It is analogous to the introduction of a power grid in the late nineteenth and early twentieth centuries. Initially, the ability to have electricity available at the flip of a switch required a significant investment in the acquisition and maintenance of infrastructure to both generate and provide the electricity. As demand increased, the infrastructure was collected into a singular power "grid." If a location was "on the grid" then power easily flowed. The cloud can be thought of as a power grid for data. If an application is "on the grid" it has access to data and services that the grid provides. The Aurora Engine is such a grid for the data we are considering in our answer to the Digging into Data challenge, and our previously discussed apps are "on" that grid. Whereas earlier efforts required that custom programming take place each time a data set was needed to provide the basis for visualization in an application, with the Aurora Engine the new application is able to simply consume data and presentation services from the cloud in a uniform way. The ability to add new apps in the future – or conversely to add new data sets to existing apps – will support considerable growth of the project beyond the term of the partnership with NEH.

Figure 4 (see Appendix) presents the high-level application architecture of the Aurora Engine and the Aurora Project web site; the architecture is divided into three tiers. At the lowest tier are the individual data sets that contribute to the overall data model for the framework. At the time of writing, the data model includes: Rail Network GIS data from 1850 to 1900; political campaign results form 1820 to 1968; railroad and Freedmen Bureau worker data; digitized newspaper data from the Richmond Daily Dispatch between November 1861 and December 1865; and geopolitical GIS historic boundary data. Each of these individual data sets is connected to higher-level application functionality through a data integration bus. Data are obtained in their source format and integrated into the framework by incorporating them with our general model for representing historic spatio-temporal data sets. This

process involves developing handlers that are capable of converting base formats for inclusion in the system. Once connected to the data bus, data can be queried in a framework native manner, allowing data from one original format to be easily related to data from another format.

At the core of the Aurora Engine is the application service tier. This tier provides software services that are connected to the data bus to provide integration, visualization, analysis, and curation functionality within the framework. These services encapsulate the workflows that researchers employ with working with large data sets, introducing data best-practices to the system while automating much of the work that is required to manually process large data sets in order to extract needed information. The application services provide the basis for the development of end-user software components using the framework. Services provided by the Aurora Engine include spatio-temporal data modelers and integrators that are capable of dynamically composing data sets components based on a location in space and a point in time. The service allow applications to provide an end user with the ability to enter a place and time and then browse the rich set of data that is applicable to that time in place. Additional services allow data to be filtered or mined based on keywords and concepts. By chaining service functionality, applications can be developed that not only allow a user to visualize entire data sets over space and time, but also allows for real-time refinement of data based on identified concepts that result from browsing data. The services, when paired with data, provide the ability to easily create advanced information discovery interfaces.

Service functionality is accessed in one of two ways in the third tier of the system; the data access tier. Applications that take advantage of the Aurora Engine can consume services and data using a set of provided Application Programmer Interfaces (APIs) built using RESTful web services.[27] The web service APIs do not constrain the implementation of applications on the Aurora Engine to any specific software platform and were designed to support a large community of application developers within the space of digital humanities with specific focus on digital history. End user access functionality is also provided in the data access tier by several web applications that serve as a

reference model for implementation of software using the Aurora Engine web services. These applications, described above, give users direct access to the data and "digging" approaches that were employed by our research team during the project and are realized by our end user interaction model, the Aurora Project. The goal of the model is three-fold, serving: 1) as a basis for presentation of our scholarly interpretation of the data that we "dug into," 2) to give the larger research community access to the same tools used to draw our own conclusions from data, and 3) to serve as a reference for development of other web applications in future efforts.

Conclusion

Since 2004, Google Inc. have digitized more than 15 million, or about 12% of the nearly 130 million books they calculate have ever been published. While statistics on the proportion of these books that are in the public domain are hard to come by, their nearest academic 'competitor', the Hathi Trust, boasts nearly 8.5 million volumes of both books and serials. Google originally digitized many of these volumes on a non-exclusive basis and 2.2 million are in the public domain. Should the Hathi Trust be able to maintain its current rate of collection development until 2020, there would be over 23 million volumes consuming more than a petabyte of digital storage, in lieu of nearly 300 miles of physical library shelving. Perhaps more significantly for historically-oriented scholars, approximately two-thirds of the current Hathi Trust public domain volumes were published between 1820 and 1920. By the end of this decade, therefore, at least 4 million public domain volumes from this period alone may be widely accessible and with improving data capture technology the total could be significantly higher still.

The 'Million books' question that motivated the Digging into Data Challenge is therefore already being superseded by the five and ten million book questions and eventually even by the hundred million book question. If Google's estimates for the number of serials volumes are added into the equation, using the Hathi Trust multipliers would suggest there are in total more than 50 billion

pages of printed text and at 250 words per page, this would yield a mere 12.5 trillion words. Yet such a total is, of course, only part of the story. To it, again from an historical perspective, should be added potentially hundreds of millions of pages of newsprint, tens of millions of maps, and innumerable illustrations and still/moving image photographic records, even before opening the immense archives of manuscript material held in the vaults of the world's major libraries, only a small fraction of which have ever been microfilmed. All these latter types of documentary sources are not simply additions to the corpus of printed works. Rather, each is of value in the accurate interpretation and contextualisation of the others, often in complex and interacting ways.

This last observation reminds us that unstructured text is much more than word frequency data, though it can, of course, be represented as such. Instead, it is riddled with ambiguity in meaning, in reference and in attribution, mediated through both a bibliographic and a spatio-temporal hierarchy of contexts. These extend from enclosing sentences and paragraphs, through to chapter, entire work and series, if appropriate, in bibliographic terms. However, the volume itself, in terms of its dates and location(s) of writing and publication also falls within historical, geographical, economic, institutional and philosophical/religious contexts. Hence, at the most trivial level, when we encounter the word 'Paris', are we dealing with Paris, France or Paris, Texas – the latter having no meaning prior to the 19th century. While such observations are largely self-evident to the educated reader of texts, the same cannot be said for a machine-learning algorithm processing a folder full of pdf files for scanned books from randomly selected dates.

However, structured datasets, such as historical census files and their historical GIS counterparts, or digitized historical map collections are far from immune to charges of selective representation of past socio-economic circumstances or past landscapes that are only partially knowable or indeed recoverable. In a recent Digital Humanities Quarterly piece, Johanna Drucker has argued strongly for replacement of the conventional term 'data' with the more technical term 'capta'. She claims that reference to 'data' aligns the investigator with a realist conception of knowledge, where

facts are ‘given’, able to be observed and recorded objectively. In contrast, use of the term ‘capta’ aligns the investigator with a constructivist position, focusing attention on the ‘taking’, even extracting of information, from a world whose characteristics, it should be noted, are not closely specified. Her main point is clear, however : "Humanistic inquiry acknowledges the situated, partial and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, not simply given as a natural representation of pre-existing fact."[28]

Over the *Longue Durée* and in the light of shifting paradigms of both humanities-based and scientific thought, it is difficult to disagree with the notion that knowledge is contingent. In the era before the advent of mass digitisation projects, this was also in an important sense unproblematic. The rank and file of scholars contented themselves with the limited spatio-temporal bounds of their studies, necessarily restricted by logistical considerations, while the self-selecting few opted for broad brush surveys of much grander swathes of space and time, picking and choosing from more detailed studies as their narrative contexts seemed to require. The much greater levels of access to huge bodies of on-line resources that are now in train look set to challenge this comfortable academic "division of labour" and the nascent field of "culturomics" may be sounding the opening notes of a new argument over the long-term nature of knowledge production and diffusion.[29]

This is particularly important in the present context, as so much of the human written and visual record becomes available in digital form. Armies of robotic book scanners have thrust before us immense new possibilities for interrogating large portions of the corpus of recorded human knowledge, life and movement in ways that were virtually unthinkable as little as a generation ago. We recognize that our project has touched on this vital issue, and we see great promise in the changing course of knowledge creation with digital techniques.

Yet, we also recognize that categories require definition for any analysis to proceed. Occupational descriptions in historical census data are a case in point. Is a machinist in Savannah, GA in 1840 the same type of worker as a machinist in Baltimore in 1910? Is the ‘simple’ term machinist

adequate to describe the range of duties performed by individual workers so labeled in the US 1880 census? Until a short time ago, possible answers to such questions could only be surmised based on essentially anecdotal evidence from literary references or small samples of data in specific locations from different sets of manuscript census schedules. Now, by contrast, as a result of the North Atlantic Population Project, the records of all 53 million individuals in the 1880 census are available for scrutiny, including both textual and coded occupational descriptions. Questions can now be asked about the extent of variations in historical enumeration quality across space, and the appropriateness of present-day occupational coding schemes or the accuracy of assignment of occupational codes to industrial sectors, to name but a few issues now open to investigation that previously were largely inaccessible to researchers.

These matters, though highly specific, have very wide ramifications now as scholars are starting to link census samples between census decades and even between different national datasets across the Atlantic. If occupation is wrongly or inadequately described in the 1880 census, when it is the key measure of socio-economic status, or the same occupational labels have different connotations in different national censuses, where does that leave studies of the economic returns to transatlantic migration, for example? The approach adopted here is to accept neither the assumption that a complex reality can be compressed into a very small number of categories, nor that it is so infinitely complex that any attempt to standardize data or search for regularities is fruitless. Once again, a critical middle ground is a more appropriate stance in our view. We use the new and larger datasets to tease out better estimates of the variability within and between groups of individuals. Further to this, it is now possible to envisage, and even commence, a very lengthy process of 'triangulation' between different datasets, be they company records, city directories or digitized census schedules, to obtain a clearer sense, for example, of how significant or useful census snapshots of occupational status actually are in the context of rapidly fluctuating nineteenth-century economic conditions.

Sufficient has been said, by way of conclusion, to indicate that the growing volume of digital

source materials should not result in the increasing suspension of our critical faculties when terabytes of data wash over us. Instead, more data requires a much greater engagement of these critical faculties, since there is more scope for detailed investigation of within and between group variability, more opportunity for comparison and integration of different types of research resources, and more need for both simple and complex analyses, data mining and data visualization.

Appendix

Figure 1: Railroad Worker App

"Central Places" from NAPP extracted 1880 data

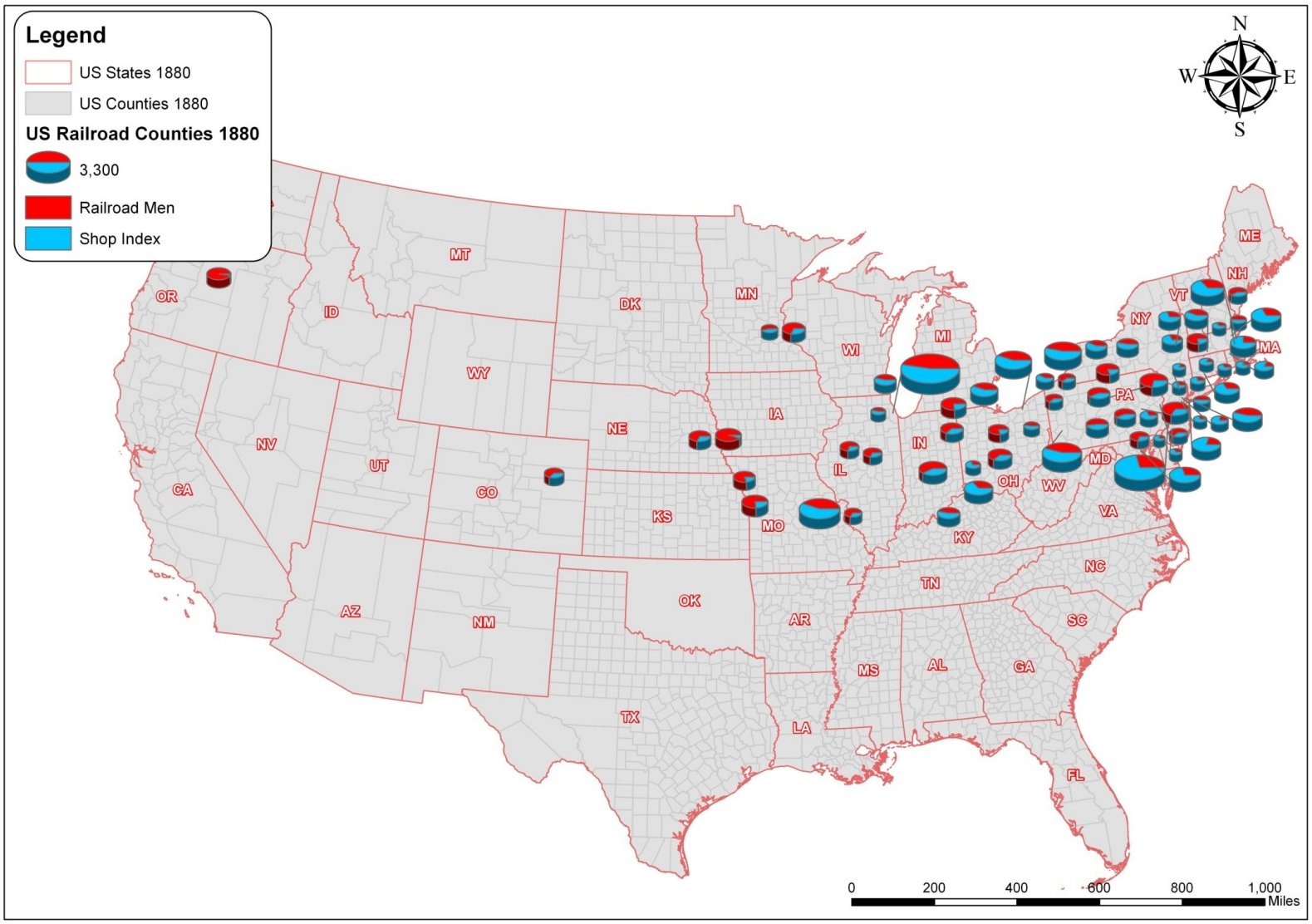

Figure 2: Network Connectivity App, the Baltimore & Ohio System

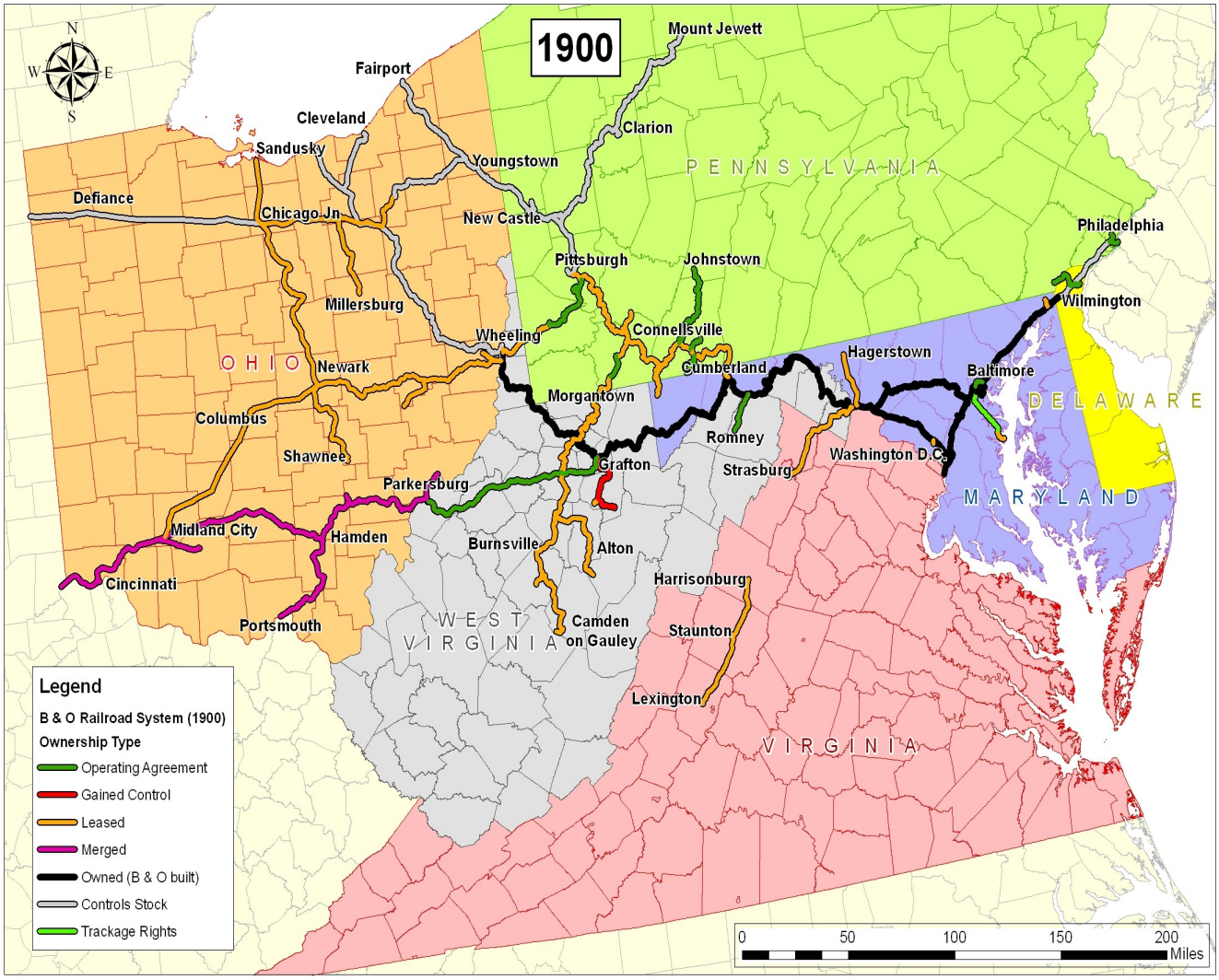

Figure 3: Aurora Engine Architecture

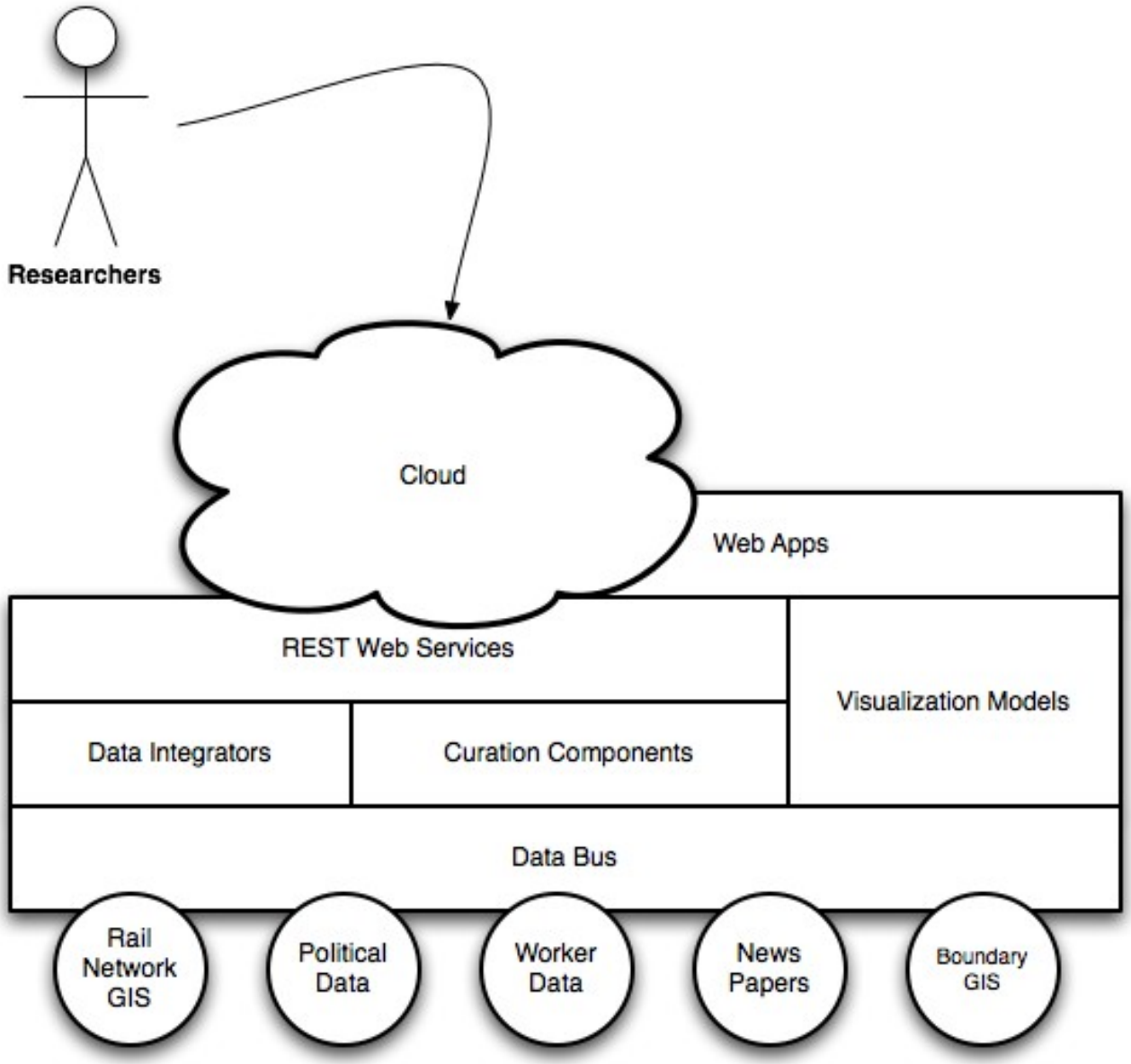

Figure 4: Sample Interface Page from the Aurora Project

[1] The classic works stressing space and time are Leo Marx, The Machine in the Garden: Technology and the Pastoral Ideal in America (Oxford: Oxford University Press, 1964), 194; and Wolfgang Schivelbusch, The Railway Journey: Industrialization and Perception of Time and Space (Sacramento: University of California Press, 1986). More recently, Richard White, "Information, Markets, and Corruption: Transcontinental Railroads in the Gilded Age," Journal of American History, Vol. 90 (2003): 19. For a detailed account of the rhetoric surrounding railroads, see Craig Miner, A Most Magnificent Machine: America Adopts the Railroad, 1825–1862 (Lawrence: University Press of Kansas, 2010). The forthcoming work by Richard White addresses the way transcontinental railroads shaped Western space and time, Richard White, Railroaded: The Transcontinentals and the Making of Modern America (New York: W. W. Norton, 2011).

[2] The social effects of the railroad bear striking resemblance to those of the Internet, especially in the ways they transform culture and in their recursive politics and economics. We are especially interested in the ways these technologies are open-ended, enabling both freedom and oppression. The railroad could lead in either direction. We have been influenced in this approach by the broader field of digital humanities and by much of the recent analysis of the digital revolution, such as Sherry Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other (New York: Basic Books, 2011), Evgeny Morozov, The Net Delusion: The Dark Side of Internet Freedom (New York: Public Affairs, 2011), and Lee Siegel, Against the Machine: Being Human in the Age of the Electronic Mob (New York: Spiegel & Grau, 2008). Richard White has also recently argued that railroads had both "software" and "hardware" systems, see http://spatialhistory.stanford.edu and Richard White, "Constructing Railroad Space," paper presented at the Social Science History Association, Chicago, Ill., November 2010. On the social construction of technologies, and for a recent account of the social meanings of technology and how people adapt technology to their use, see David Edgerton, The Shock of the Old: Technology in Global History Since 1900 (New York: Oxford University Press, 2006). Edgerton emphasizes a history of "technology in use" rather than invention and the persistence of old technologies among the modern. He calls the tendency to overemphasize the impact of technology "futurism." Here, Edgerton's view is especially relevant because with railroads the question is how people adjusted to them, adapted, and came to terms with their use and meaning. This is predominately a cultural and social question. Other important works focused on this question include: Carolyn Marvin, When Old Technologies Were New: Thinking About Electric Communication in the Nineteenth Century (New York: Oxford University Press, 1988), esp. 193–209; and David Nye, Technology Matters: Questions to Live With (Cambridge: MIT Press, 2006), esp. 46–47. Nye emphasizes that technology is not deterministic and is "unpredictable," often with "no immediate impact."

[3] See 1896 (http://projects.vassar.edu/1896/1896home.html); The Spatial History Project (http://spatialhistory.stanford.edu); Virtual Shanghai (http://www.virtualshanghai.net/). The American Historical Review pieces include: Robert Darnton, "An Early Information Society: News and the Media in Eighteenth Century Paris" (http://www2.vcdh.virginia.edu/AHR/), Philip J. Ethington, "Los Angeles and the Problem of Historical Knowledge" (http://cwis.usc.edu/dept/LAS/history/historylab/LAPUHK/index.html), William G. Thomas III and Edward L. Ayers, "The Differences Slavery Made: A Close Analysis of Two American Communities" (http://www2.vcdh.virginia.edu/AHR/), Jack Censer and Lynn Hunt, "Imagining the French Revolution: Depictions of the French Revolutionary Crowd" (http://chnm.gmu.edu/revolution/imaging/home.html). All are available at: http://www.indiana.edu/~ahrweb/projects.html.

[4] Fernand Bruadel, *On History* (Chicago: University of Chicago Press, 1980): 76. Quoted in Keith Thomas, "A Highly Paradoxical Historian," *New York Review of Books*, April 12, 2007, p. 56.

[5] Edward L. Ayers, "Turning Toward Place, Space, and Time," and Voting America <http://www.votingamerica.org>, Digital Scholar Lab, University of Richmond. Only rarely does a historian choose a narrative form that explicitly reveals these complex spatial and temporal relationships. Hans Gumbrecht's In 1926: Living at the Edge of Time offers a useful example. Gumbrecht wanted to reveal simultaneity and "the existence of a 'web' or 'field' of not only discursive realities that strongly shaped the behavior and interactions of 1926." Because he is concerned with representing simultaneity--rather than sequentiality-- he is not especially concerned with subjective or collective agency. Instead, his aim is to make "present a historical environment of which we know (nothing more than) that it existed in some places during the year 1926." So, Gumbrecht's history uses categories and cross-indexing. Its narrative is woven and interlaced, exposing the associations Gumbrecht wants us to understand and thrusting the reader right into the middle of them. Hans Ulrich Gumbrecht, In 1926: Living at the Edge of Time (Cambridge: Harvard University Press, 1997): xi.

[6] For the view of history as climate, Fernand Braudel, Civilization and Capitalism: Volume 3 The Perspective of the World (New York: Harper and Row, 1979): 618; as drama, see Rhys Isaac, The Transformation of Virginia (New York: W. W. Norton): appendix.

[7] Sherry Ortner, Anthropology and Social Theory: Culture, Power, and the Acting Subject (Durham: Duke University Press, 2006); Anthony Giddens, The Consequences of Modernity (Cambridge: Polity, 1990), The Constitution of Society, Outline of the Theory of Structuration (Berkeley: University of California Press, 1984); Anthony Giddens and Christopher Pierson, Conversations with Anthony Giddens: Making Sense of Modernity (Cambridge: Polity Press, 1998); for a useful overview of this literature, see William H. Sewell, Logics of History: Social Theory and Social

Transformation (Chicago: University of Chicago Press, 2005); Pierre Bourdieu, The Logic of Practice (Cambridge: Polity Press, 1990) and Outline of a Theory of Practice (Cambridge: Cambridge University Press, 1977); Raymond Williams, Culture and Society, 1780-1950 (New York: Columbia University Press, 1983).

[8] Our concerns and approaches have been influenced throughout especially by "actor network theory," as well as the literature on the social construction of space. Bruno Latour's views on the modern "constitution" with its separation of Nature and Society, on the importance nonhuman objects (what he calls "hybrids") as actors, on the idea of a "sociology of associations," and on the mediation among actors in society have been especially important in our thinking about the role of the railroad and the war in Southern and American society. Indeed, we seek to trace the ways the railroad served as a hybrid agent, mediating any number of social, political, and economic associations. Our goal throughout is to follow the railroad's associations, map them, and place them in relation to one another. See Bruno Latour, Reassembling the Social: An Introduction to Actor Network Theory (Oxford: Oxford University Press, 2005) and We Have Never Been Modern (Cambridge: Harvard University Press, 1993). The works on space and modernity have also influenced the overall emphasis in this project, including: J. Nicholas Entrikin, The Betweenness of Place: Towards a Geography of Modernity (Macmillan, 1991), esp. 27-59; Allan Pred, Making Histories and Constructing Human Geographies: The Local Transformation of Practice, Power Relations, and Consciousness (Boulder: Westview Press, 1990), esp. 126-170.

[9] The term "second nature" is drawn from William Cronon's Nature's Metropolis: Chicago and the Great West (New York: W. W. Norton, 1991), xix, 62–78, and in his edited volume, Uncommon Ground: Rethinking the Human Place in Nature (New York: W. W. Norton, 1995). Cronon, an environmental historian, is interested in the way rails, commodity markets, commercial lending, and other systems altered the natural landscape in the 1870s and 1880s, especially around the city of Chicago, but in the Civil War the distinction is equally significant. Cronon calls second nature "the artificial nature that people erect atop first nature" and recognizes the ambiguity in the terms and the "complex mingling of the two."

[10] Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences (2006). (http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf)

[11] For recent definitions and explorations of the methodology of digital history, see Douglas Seefeldt and William G. Thomas, "What is Digital History: A Look at Some Exemplar Projects," Perspectives on History (May 2009); Daniel J. Cohen and Roy Rosenzweig, Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web (Philadelphia: University of Pennsylvania Press, 2006); Orville Vernon Burton, ed., Computing in the Social Sciences and Humanities (Urbana and Chicago: University of Illinois Press, 2002); David J. Staley, Computers, Visualization, and History: How New Technology will Transform our Understanding of the Past (Armonk, N.Y.: M. E. Sharpe, 2003); Edward L. Ayers, "The Pasts and Futures of Digital History," Virginia Center for Digital History, 1999 <http://www.vcdh.virginia.edu/PastsFutures.html>; Orville Vernon Burton, "American Digital History," Social Science Computer Review 23.1 (2005): 206-220; Daniel J. Cohen, Michael Frisch, Patrick Gallagher, Steven Mintz, Kirsten Sword, Amy Murrell Taylor, William G. Thomas III, and William J. Turkel, "Interchange: The Promise of Digital History," Journal of American History Vol. 95, No. 2 (2008) <http://www.journalofamericanhistory.org/issues/952/interchange/index.html>.Digital history as a scholarly practice bears a close relationship to the wider field of Digital Humanities. Some of the key influences on our thinking about interfaces, texts, and narratives have been: Jerome McGann, Radiant Textuality (New York: Palgrave, 2001); Johanna Drucker, Speclab: Digital Aesthetics and Projects in Speculative Computing (Chicago and London: University of Chicago Press, 2009); Lev Manovich, The Language of New Media (Cambridge: MIT Press, 2001); Susan Schreibman, Ray Siemens, and John Unsworth, A Companion to Digital Humanities (Malden, MA: Blackwell, 2004); Anthony Grafton, The Footnote: A Curious History (Cambridge: Harvard University Press, 1997); John Seeley Brown and Paul Duquid, The Social Life of Information (Cambridge: Harvard Business Press, 2000); Espen Aarseth, Cybertext: Perspectives on Ergodic Literature (Baltimore: Johns Hopkins Press, 1997); and Janet Murray, Hamlet on the Holodeck: the Future of Narrative in Cyberspace (New York: Free Press, 1997).

[12] For a recent application of computation on large data sets, see Jean-Baptiste Michel, Erez Lieberman Aiden, et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," Science Vol. 331 (14 January 2011). See also for example, Franco Moretti, Graphs, Maps, and Trees (New York: Verso, 2005). On the spatial turn in history and humanities, see David J. Bodenhamer, John Corrigan, and Trevor M. Harris, eds., The Spatial Humanities: GIS and the Future of Humanities Scholarship (Bloomington: Indiana University Press, 2010).

[13] See especially, Charles Henry and Kathlin Smith, "Ghostlier Demarcations: Large-Scale Text Digitization Projects and Their Utility for Contemporary Humanities Scholarship," in Charles Henry, et al. "The Idea of Order: Transforming Research Collections for 21st Century Scholarship," (Washington, D.C.: Council on Library and Information Resources, June 2010). Robert Darnton, The Case for Books: Past, Present, and Future (New York: Public Affairs, 2009).

[14] The Apps were distributed on August 16, 2010 to our Digging into Data advisory group and partners, including John Lutz, Sherry Olson, Anne Bretagnolle, Ian Gregory, Richard White, and Anne Knowles. Each participant responded with detailed suggestions and comments on the research questions.

[15] The most detailed analysis of railroad workers and social mobility remains Walter Licht, Working for the Railroad: The Organization of Work in the Nineteenth Century (Princeton: Princeton University Press, 1983), 77. In one of his more important conclusions Licht found considerable turnover--over 50 % in six months and only 25 % of workers stayed with a railroad company on the payroll for two years. Also see Peter Way, Common Labor: Workers and the Digging of North American Canals, 1780-1860 (New York: Cambridge University Press, 1993) and David R. Meyer, Networked Machinists: High-Technology Industries in Antebellum America (Baltimore: Johns Hopkins University Press, 2006), 8-17 on the formation of sub-regional and interregional networks, and146-177 on the development of railroad locomotive machine shops. Meyer argues that the machinists formed their networks early in the 1830s and 40s (p. 20) and that in the East they numbered "perhaps two to three thousand machinists between 1820 and 1840" (p. 17). Meyer argues that the networks shift to the Midwest in the 1860s (p. 280).

[16] Stromquist, A Generation of Boomers, 238. Paul Black compiled a computerized database file of all 8,092 discharged employees on the Burlington's list, but the file only remains in the CBQ Collection at the Newberry in hard copy print. See Paul V. Black, "Experiment in Bureaucratic Centralization: Employee Blacklisting on the Burlington Railroad, 1877-1892," Business History Review Vol. 51 No. 4 (Winter 1977): 444-459.

[17] Stromquist, A Generation of Boomers, 220-222.

[18] Stromquist, A Generation of Boomers, 160-161 and footnote 51 on p. 307. Stromquist's focus on "market towns" such as Burlington and "railroad towns" such as Crestone accounted for differences in the ways smaller communities responded to strikes and changing class relationships. Stromquist questioned Herbert Gutman's interpretation of industrial strikes as defined by a pattern of community support for working-class members who took on "outside" corporations. Gutman emphasized the small-town, face to face relationships that sustained these "organic" communities. He contrasted these communities with the large metropolises of New York and Chicago. Stromquist considered smaller railroad towns, however, subject to change and variation. See Stromquist, A Generation of Boomers, 147-8. He concluded that the railroad places were diverse and developed differently, the relationship of classes was not static, community solidarity not necessarily a given. High turnover of population did not weaken solidarity if labor was scarce (as it was for the first phase of development), but as labor supply expanded and railroads turned to involuntary movement of labor, social conflict resulted. The generation of boomers was over--their solidarity, even as they moved rapidly across space, weakened not by the movement itself but by the changing labor conditions--essentially the forced lack of labor scarcity which had formed the generation of boomers.

[19] W. F. Merrill to Paul Morton, June 23, 1888. CBQ Collection, Newberry Library.

[20] G. M. Hohl to W. F. Merrill, June 14, 1888. CBQ Collection, Newberry Library.

[21] List of Fireman and Hostelers making 1608 applications, to G. W. Rhodes, February 11, 1889. CBQ Collection, Newberry Library. See Sheldon Stromquist, A Generation of Boomers: The Pattern of Railroad Labor Conflict in Nineteenth-Century America (Urbana and Chicago: University of Illinois Press), 238. No recent studies have examined worker mobility or migration in the Great Plains.

[22] R. G. Healey, "A Full-Scale Implementation of the NAPP 1880 U.S. Census Dataset Using Dimensional Modeling and Data-Warehousing Technology," Historical Methods (2011 in press).

[23] Charles O. Paullin and John K. Wright, Atlas of the Historical Geography of the United States (Carnegie Institute of Washington and the American Geographical Society of New York, 1932), 138.

[24] The classic studies of the post-emancipation South include: Roger L. Ransom and Richard Sutch, One Kind of Freedom: The Economic Consequences of Emancipation (New York: Cambridge University Press, 1977); Gavin Wright, The Political Economy of the Cotton South: Households, Markets, and Wealth in the Nineteenth Century (New York: W. W. Norton, 1978), esp. 158-184; Robert Higgs, Competition and Coercion: Blacks in the American Economy, 1865-1914 (New York: Cambridge University Press, 1977); Leon F. Litwack, Been in the Storm So Long: The Aftermath of Slavery (New York: Alfred A. Knopf, 1979); Gavin Wright, Old South, New South: Revolutions in the Southern Economy since the Civil War (New York: Basic Books, 1986); Stephen Hahn, A Nation Under Our Feet: Black Political Struggles in the Rural South from Slavery to the Great Migration (Cambridge, Mass.: Belknap Press, 2003). For a new look at this question, especially the question of scale and the process of emancipation, see Edward L. Ayers and Scott Nesbit, "Seeing Emancipation: Scale and Freedom in the American South," The Journal of the Civil War Era Vol. 1, No. 1 (March 2011), 3-24. Also, for a promising new analysis of freedmen and landholding, see Melinda Miller, "Using Quantification to Establish Causation: Forty Acres and a Mule Would Have Made a Difference," paper presented at the Organization of American Historians, March 2011, Houston, TX.

[25] On the concept of gateways, see Carlton J. Corliss, Main Line of Mid-America: The Story of the Illinois Central (New York: Creative Age Press, 1950), 137. On the Freedmen's Bureau data, we have used the records held at National Archives, Record Group 105, Records of Assistant Commissioners, Chattanooga M1911 Roll 15, Louisville M1904 Roll 123, Camp Nelson M1904 Roll 64-66, Petersburg M1913 Roll 160, Alexandria M1913 Roll 52.

[26] The literature on railroads and conservation is limited but growing in significance. An essential starting point is Richard Orsi, Sunset Limited: The Southern Pacific Railroad and the Development of the American West, 1850-1930 (Sacramento: University of California Press, 2005). For a Southern account, see Mart A. Stewart, 'What Nature

Suffers to Groe': Life, Labor, and Landscape on the Georgia Coast, 1680-1920 (Athens: University of Georgia Press, 1996). See the forthcoming work by Sean M. Kammer, "'The Railroads Must Have Ties': Edward H. Harriman and Forest Conservation, 1901-1908," Western Legal History Vol. 23, No. 1 (forthcoming). One of the earliest and most important works remains Sherry H. Olson, The Depletion Myth: A History of Railroad Use of Timber (Cambridge: Harvard University Press, 1971). The story of timber and timber resource extraction alone is complex--see David A. Clary, Timber and the Forest Service (Lawrence: University Press of Kansas, 1986) Richard C. Overton, Burlington West: A Colonization History of the Burlington Railroad (Cambridge: Harvard University Press, 1941), 245 on timber depredations and poaching, p. 294 on the Burlington policy not to grant ten year contracts for timber lands because settlers had stripped the timber and defaulted on the land. See "Pre Emption Circular" Burlington & Missouri River Railroad Co., "conditions of pre-emption," November 15, 1870. Burlington & Missouri River Railroad in Nebraska Land Department Applications, Vol. 1, Nebraska State Historical Society, Record Group 3508, microfilm 17716. On the widespread problem of timber poaching (a concern to the land grant railroads and the government), see The Atchison Daily Globe, November 21, 1894, "Wanton Destruction," The Rocky Mountain News, February 15, 1896, "Work of Timber Thieves," The Daily Evening Bulletin, February 23, 1883, "Instructions Concerning Timber Lands on Railroad Grants." Railroads, too, were guilty of promoting timber poaching on government lands and some ranged far and wide under the loose clauses granting them the right to take timber for construction and operation. See also William H. Sellew, Railway Maintenance Engineering with notes on Construction (New York: D. Van Nostrand Company, 1919), 89 on railroad "tree plantations" and their lack of success, as well as the recommendation to railroads to cultivate native trees on their cut-over forests. Sellew also notes that "several roads hold timber lands, which they have placed under management with a view to providing a source of tie supply." There were Union Pacific experiments in tree planting on the plains to prove that trees might grow along the line, and the Missouri River, Fort Scott & Gulf Railroad conducted a widely reported experiment in a railroad managed "tree plantation." See "Tree Planting for Railroads" by John A. Warder in Forestry: Report of Delegation appointed to attend the American Forestry Congress (Toronto, 1882), 119. Samuel P. Hays, Conservation and the Gospel of Efficiency: The Progressive Conservation Movement, 1890-1920 (Harvard University Press, 1959). The most recent study of forests and forest policy is Char Miller, ed., American Forests: Nature, Culture, and Politics (University Press of Kansas, 1997). In "Timber Users, Timber Savers: The Homstake Mining Company and the First Regulated Timber Harvest," Richmond L. Clow demonstrates that industry defined conservation practices as much as the government. Mining companies and railroads became leading institutional players in the scientific management of forests between 1900 and 1910. Roy M. Robbins, Our Landed Heritage: The Public Domain, 1776-1970 (Lincoln: University of Nebraska Press, 1942), 339-340. Railroads, Robbins notes, "became aggressive advocates of the forest reserve policy" and tried "to juggle the government out of its valuable timber."

[27] Roy Thomas Fielding, Architectural Styles and the Design of Network-based Software Architectures (Ph.D. Dissertation, University of California, Irvine, 2000).

[28] Johanna Drucker, "Humanities Approaches to Graphical Display," Digital Humanities Quarterly Vol. 5, No. 1 (Winter 2011) http://digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

[29] Jean-Baptiste Michel, Erez Lieberman Aiden, et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," Science Vol. 331 (14 January 2011). See www.culturomics.org.